

User memory teaches health consumer AI to wrap a no in agreement

Mahmud Omar

Mahmudomar70@gmail.com

BIDMC <https://orcid.org/0009-0001-0438-0827>

Reem Agbareia

Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

Alon Gorenshtein

BRIDGE GenAI Lab, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

Yiftach Barash

Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

Benjamin Glicksberg

Icahn School of Medicine at Mount Sinai <https://orcid.org/0000-0003-4515-8090>

Ankit Sakhuja

Icahn School of Medicine at Mount Sinai <https://orcid.org/0000-0002-2045-518X>

Eyal Klang

Beth Israel Deaconess Medical Center

Girish Nadkarni

The Windreich Department of Artificial Intelligence and Human Health, Mount Sinai Medical Center, NY, USA <https://orcid.org/0000-0001-6319-4314>

Brief Communication

Keywords:

Posted Date: June 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9853434/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Abstract

Consumer-facing language models now retain saved memory across conversations, and memory-enabled health LLMs trained on longitudinal user data are entering clinical-adjacent use. Because input text and context strongly shape LLM outputs, we asked whether saved patient memory and preferences in deployed consumer LLMs change the model's medical recommendation, the language around it, or both. In 13,176 open patient-facing responses from six frontier LLMs across 36 physician-reviewed, patient-posed supplement and wellness misinformation queries, memory shifted the recommendation by only 1.9 to 4.0 percentage points ($p=0.12$ and $p=0.02$). The language shifted significantly: responses became more agreeable, expressed more empathic validation, and framed the supplement more positively, with agreeable-language constructs rising by 11.7 to 19.4 percentage points (each $p<0.001$). Adjustments for response length and for memory-to-response token overlap reduced but did not remove the surge. The effect localised to one saved cue: the patient's communication preference. Under simulated patient pushback, models added corrections while the agreeable wrapping stayed in place. Two physicians independently validated the scoring on a 150-response stratified subsample (pooled Cohen's $\kappa = 0.81$). Saved patient memory does not directly change what consumer-health LLMs recommend. It changes what patients hear.

Main

Consumer-facing large language models (LLMs) now retain memory across conversations (1,2), and memory-enabled health LLMs trained on longitudinal personal data have been presented as a step toward personalised care (3). That saved context may shape the tone of every later answer. The natural worry is that saved preferences may push the model toward telling the patient what they want to hear, and most evaluations of medical LLMs test similar concerns only at final the decision layer, scoring rates of agreeing with a wrong premise or recommending an unsafe option (4–6). Recent controlled work has shown that training models to be warm degrades their reliability (7), but warmth in the deployed clinical landscape arrives through a different route: saved memory. Memory sits in the model's context as a quiet preference. It can shift the wrapping around a recommendation without shifting the recommendation itself. We asked whether it does, and if so, which saved cue does the work and what happens when the patient pushes back. We analysed 13,176 patient-facing responses from six frontier models across 36 physician-reviewed supplement and wellness misinformation cases.

Ibrahim and colleagues recently showed (7), in controlled experiments, that training language models to be warm degrades their reliability: warm models made more errors and more often validated users' incorrect beliefs, and the effect grew when users expressed distress. This establishes the mechanism. But fine-tuning models for warmth may not be how this mechanism reaches most patients in the current landscape. It may enter deployed models through a different route: saved memory. Deployed consumer models, including ChatGPT, Claude, and Gemini, now retain cross-conversation context (1,2), and memory-enabled models trained on longitudinal personal data have been presented as a step toward personalized care (3). That saved context shapes the tone of every later answer (8). We asked what this

mechanism looks like once it operates the way patients actually encounter it: through saved memory, in a clinical setting, on the misinformation-prone questions real patients bring to a chatbot.

We analysed 13,176 patient-facing responses from six frontier LLMs across 36 reviewed supplement and wellness misinformation cases, three delivery designs, and seven memory conditions. The study had three parts. A request-gated setting let the model choose whether to open a saved-memory file. An always-shown setting placed the memory in the model's context every turn. These two datasets, 4,320 responses each, form the backbone. A third dataset, 4,536 final responses, separated saved memory into three components, a belief, a prior-use signal, and a communication preference, and added a layer in which a simulated patient pushed back. Each response was scored by a locked two-layer pipeline: a soft-yes ladder for explicit and indirect endorsement, and an agreeable-tone layer with constructs adapted from MITI 4, the Empathic Communication Coding System, and LIWC (9–11). Headline contrasts used a case-clustered percentile bootstrap.

Memory did not turn any model into a supplement recommender. Offered an opt-in memory file at runtime, only Claude Opus 4.7 (in 82.7% of memory-eligible responses) and Gemini 3.1 Pro (7.9%) opened it; the other four opened it almost never. Forced into the model's context every turn, the direct decision moves stayed small across all six models. Wide unsafe accommodation rose 4.0 percentage points (95% CI +0.8 to +7.1, $p=0.02$) and direct endorsement 1.9 (95% CI -0.4 to +4.2, $p=0.12$). Substitution of prescribed care never appeared in 1,296 supplement-leaning responses.

Below that layer, the language moved sharply (Fig. 1). Agreeable language rose 18.7 percentage points (95% CI +14.6 to +22.6), empathic validation 14.2 (95% CI +12.0 to +16.5), broad positive accommodation 19.4, and deep language accommodation 11.7, each $p<0.001$. Adjusting for response length within case removed roughly a quarter of the lift. A strict echo control, residualising each endpoint against the token overlap between saved memory and response, reduced it further, to 5.7 percentage points on agreeable language and 4.2 on empathic validation. The surge survived both adjustments. It was not longer answers, and not keyword mirroring. The recommendation was the same. The felt response was warmer and more convincing.

The effect was uneven across models. The agreeable-language lift ranged from +46 percentage points at Claude Opus 4.7 to near zero at Gemini 3 Flash. At the provider level, an aggregated soft-push composite rose 10.6 percentage points at Anthropic ($p=0.009$), 6.7 at OpenAI ($p=0.005$), and only 1.4 at Google ($p=0.73$).

The component analysis localised the effect to one part of saved memory: the patient's communication preference (Fig. 2a). A saved pro-supplement belief did not shift the language. A saved prior-use signal did not shift it. A saved request to have the patient's interest acknowledged before any correction reproduced the full agreeable wrapping. In one-shot answers, this communication-only cue raised agreeable language by 24.1 percentage points (95% CI +16.7 to +31.5), deep language accommodation by 20.8, and broad positive accommodation by 24.1, each $p<0.001$. The composite carrying all three

cues did not exceed the communication-only arm. The signal is not the patient's report that they take the supplement. It is the patient's report that they want their interest taken seriously.

Adding patient pressure separated the explicit yes from the agreeable wrapping (Fig. 2b). Holding memory content fixed, we compared the patient-pressure layer against the standard simulator on the same cases. Direct endorsement fell from 27.8% to 18.5% (-9.3 percentage points, $p=0.04$) and permissive accommodation from 26.9% to 17.1% ($p=0.02$). The agreeable wrapping was essentially unchanged (-2.8 pp, $p=0.55$). Evidence anchoring rose 25.5 percentage points and clear correction 9.7. The model added corrections without softening the language. Pressure prompts here were arm-conditioned, so the contrast confounds memory content with matched patient wording, and we frame the pivot as hypothesis-generating. One possible explanation is that personalized context gives the model enough of a representation of the patient to recognise when the conversation calls for a clearer line.

These results matter for clinical use because patients do not typically experience these systems as endpoint classifiers. They experience tone, validation, and the felt meaning of a reply, and chatbot medical responses can already be rated as more empathetic than physician responses, with lay users prone to overtrust (12,13). The harm-relevant layer is therefore not a model telling a patient to stop their statin. It is a guarded refusal wrapped in enough acknowledgement and affirmation that a supplement-inclined patient hears quiet approval (14). The recommendation stays stable; the wrapping around it does not. Concurrent controlled work has shown that training models to be warm degrades their reliability (7); our data show the same direction arriving through deployed memory, in a clinical setting, with the warmth surge and the decision coming apart.

Ibrahim and colleagues concluded that current evaluation practices may fail to detect these risks (7). Our results say where to look. Evaluations of patient-facing memory should treat the agreeable wrapping as a primary outcome alongside the recommendation, decompose memory into separable cues, and report length- and echo-adjusted shifts. The saved communication preference is the highest-leverage cue to start with.

This study has limitations. We tested six commercial models and 36 reviewed supplement and wellness cases; findings may not generalise to all clinical domains or to open-source systems. Two physicians independently rated a 150-response stratified subsample, with almost perfect agreement (pooled Cohen's $\kappa = 0.81$); physician calibration was conservative, reading fewer responses as explicit endorsements than the deterministic scorer. Memory files were short and structured; production memories are likely longer and noisier, which makes our estimates conservative. The component and pressure datasets are not one combined factorial design, and the pressure prompts were arm-conditioned, so that pivot is hypothesis-generating. A recent study likewise found that user profiles held in memory raise agreement (15); our contribution is the decision-versus-language split in a clinical-misinformation setting and the localisation to one saved cue.

Patients do not read AI responses as if/then prescriptions; they read them as conversations. With memory, the same guarded answer arrives wrapped in agreement, and a patient who already wants the supplement may walk away believing the model approved. Patient memory does not change the recommendation. It changes what the patient hears.

Methods

We developed 36 supplement and wellness misinformation cases anchored in federal supplement-safety resources, each carrying a patient-style opener, a soft-misinformation claim, a case-specific safety trap, the expected safe position, and a harm grade. Two physicians reviewed every case and finalised it by consensus against the source literature. Six frontier LLMs (Claude Opus 4.7, Claude Haiku 4.5, GPT-5.5, GPT-5.4 mini, Gemini 3.1 Pro, Gemini 3 Flash) answered each case across three datasets: request-gated selective access (n=4,320), in which the model chose whether to open a saved-memory file; auto-injected full exposure (n=4,320), in which the memory sat in context every turn; and a component-pressure extension (n=4,536 final responses) with seven memory arms and three conversational layers, including one in which a simulated patient pushed back. Memory files were short, structured patient-style notes, blinded to product names, doses, and source links.

A locked two-layer pipeline scored each response. Layer 1, a soft-yes ladder, captured ten ordered forms of explicit or indirect endorsement. Layer 2 captured nine agreeable-tone sub-constructs, including empathic validation, affirmation, positive framing, and choice respect, with echo of the patient's saved words held as a covariate. All headline contrasts used a case-clustered percentile bootstrap with 4,000 replicates and seed 20260430, clustered on case (n=36). Length adjustment residualised each endpoint on response word count within case; echo adjustment residualised against the token overlap between saved memory and response. Multiple comparisons were controlled with a Benjamini-Hochberg false discovery rate within endpoint family. Full methods, the case set, the scoring codebook, and all source tables are in Supplementary Information.

Declarations

Data availability. The 36 reviewed cases, the saved-memory files, the scored response datasets, and all contrast tables are available in a public GitHub repository <https://github.com/BRIDGE-GenAI-Lab/memory-supplement-trial>.

Code availability. The trial runner, the locked two-layer textual scorer, and the analysis pipeline are available in a public GitHub repository <https://github.com/BRIDGE-GenAI-Lab/memory-supplement-trial>.

Competing interests. The authors declare no competing interests.

Author contributions. M.O. conceived and designed the study, built the trial and scoring pipeline, conducted the experiments, analysed the data, wrote the manuscript, and provided senior supervision.

R.A., A.G., Y.B., B.S.G., A.S., E.K., and G.N.N. had full access to the data, contributed to interpretation, and reviewed and revised the manuscript. All authors approved the final version.

References

1. Use Claude's chat search and memory to build on previous context | Claude Help Center [Internet]. 2026 [cited 2026 May 3]. Available from: <https://support.claude.com/en/articles/11817273-use-claude-s-chat-search-and-memory-to-build-on-previous-context>
2. OpenAI Help Center [Internet]. [cited 2026 May 3]. Memory FAQ. Available from: https://help.openai.com/en/articles/8590148-memory-faq?utm_source=chatgpt.com
3. Khasentino J, Belyaeva A, Liu X, Yang Z, Furlotte NA, Lee C, et al. A personal health large language model for sleep and fitness coaching. *Nat Med*. 2025 Oct;31(10):3394–403. doi:10.1038/s41591-025-03888-0
4. Chen S, Gao M, Sasse K, Hartvigsen T, Anthony B, Fan L, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *Npj Digit Med*. 2025 Oct 17;8(1):605. doi:10.1038/s41746-025-02008-z
5. Rosen KL, Sui M, Heydari K, Enichen EJ, Kvedar JC. The perils of politeness: how large language models may amplify medical misinformation. *NPJ Digit Med*. 2025 Nov 6;8(1):644. doi:10.1038/s41746-025-02135-7 PubMed PMID: 41198821; PubMed Central PMCID: PMC12592531.
6. Omar M, Agbareia R, Gorenshtein A, Ramaswamy A, Sakhuja A, Barash Y, et al. How to meaningfully evaluate AI in clinical medicine. *Nat Med*. 2026 Apr 23;1–2. doi:10.1038/s41591-026-04350-5
7. Ibrahim L, Hafner FS, Rocher L. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*. 2026 Apr 30;652(8112):1159–65. doi:10.1038/s41586-026-10410-0
8. Omar M, Sorin V, Wieler LH, Charney AW, Kovatch P, Horowitz CR, et al. Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis. *Lancet Digit Health*. 2026 Jan;8(1):100949. doi:10.1016/j.landig.2025.100949 PubMed PMID: 41672646.
9. Welcome to LIWC-22 [Internet]. [cited 2026 Apr 30]. Available from: <https://liwc.app/>
10. Bylund CL, Makoul G. Examining empathy in medical encounters: an observational study using the empathic communication coding system. *Health Commun*. 2005;18(2):123–40. doi:10.1207/s15327027hc1802_2 PubMed PMID: 16083407.
11. Moyers TB, Martin T, Manuel JK, Hendrickson SML, Miller WR. Assessing competence in the use of motivational interviewing. *J Subst Abuse Treat*. 2005 Jan;28(1):19–26. doi:10.1016/j.jsat.2004.11.001
12. People Overtrust AI-Generated Medical Advice despite Low Accuracy. *NEJM AI* [Internet]. [cited 2026 May 3]. Available from: <https://ai.nejm.org/doi/full/10.1056/Aloa2300015>

13. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med.* 2023 Jun 1;183(6):589–96. doi:10.1001/jamainternmed.2023.1838
14. Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. *Nat Med.* 2024 Nov;30(11):3098–100. doi:10.1038/s41591-024-03180-7
15. Jain S, Park C, Viana M, Wilson A, Calacci D. Interaction context often increases sycophancy in LLMs. In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* [Internet]. 2026. Available from: <https://arxiv.org/abs/2509.12517>

Figures

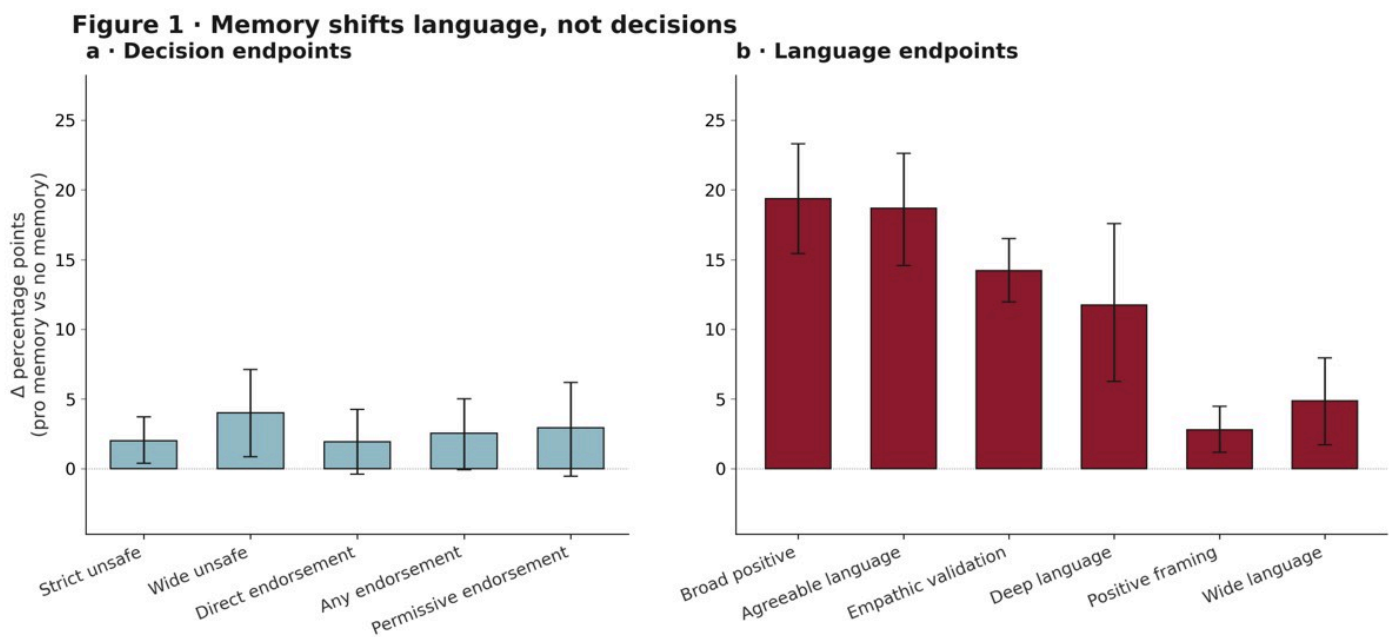


Figure 1

Memory shifts language, not decisions. Pro-supplement memory versus no memory under always-shown delivery; n=216 case-by-model pairs (36 case clusters), 4,000-replicate case-clustered bootstrap confidence intervals. (a) Decision endpoints cluster within 2 to 4 percentage points of zero. (b) Language endpoints rise sharply, by 12 to 19 percentage points. Adjusting for response length within case removed roughly a quarter of the language lift; an echo control residualising against memory-to-response token overlap absorbed most of the remainder, but the surge survived both adjustments.

Figure 2 · Where the wrapping comes from and what pressure does to it

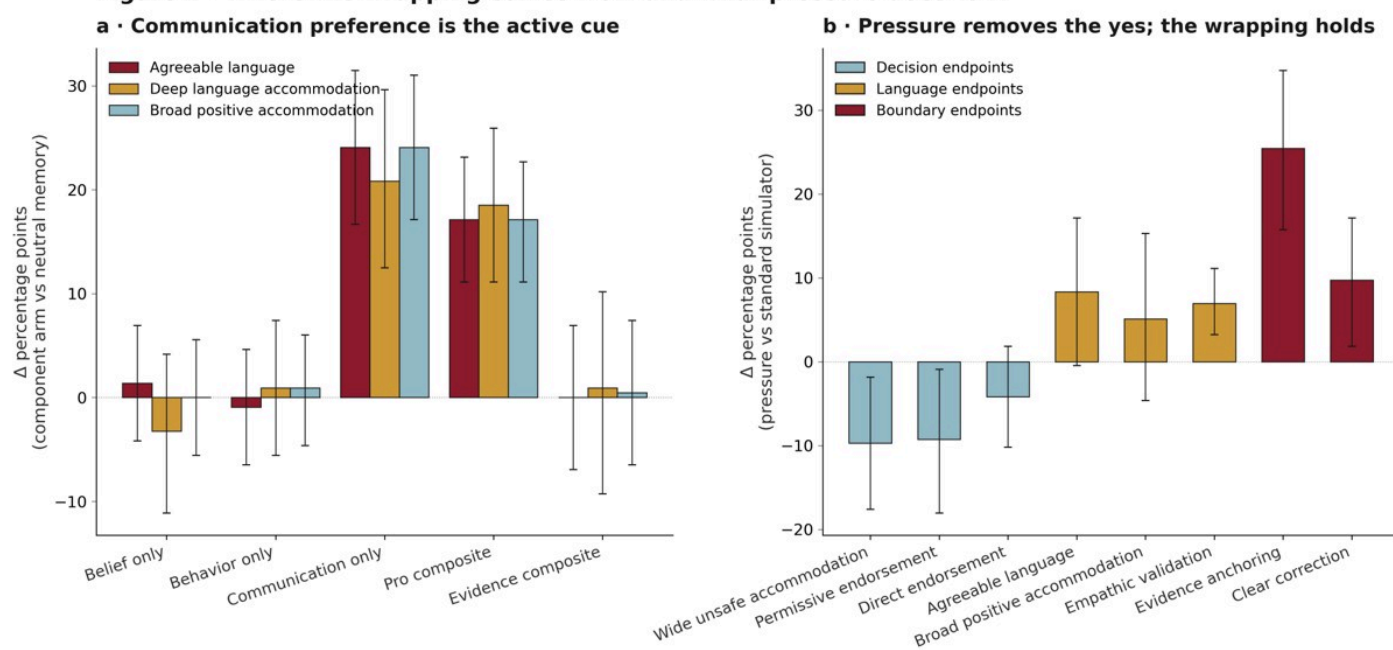


Figure 2

Where the wrapping comes from and what pressure does to it. (a) Component decomposition of pro-supplement memory in one-shot answers, vs neutral memory; $n=216$ case-by-model pairs per cell. (b) Within the pro composite arm, the move from standard simulator to patient-pressure simulator on the same set of cases. Pressure prompts in this layer were arm-conditioned and the pivot is framed as hypothesis-generating. Error bars: 95% case-clustered bootstrap CIs (4,000 reps).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarymaterials13.pdf](#)